

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **06274548 A**

(43) Date of publication of application: **30 . 09 . 94**

(51) Int. Cl.

G06F 15/40
G06F 15/38

(21) Application number: **05061641**

(22) Date of filing: **22 . 03 . 93**

(71) Applicant: **A T R JIDO HONYAKU DENWA**
KENKYUSHO:KK

(72) Inventor: **OI KOZO**
SUMIDA EIICHIRO
IIDA HITOSHI

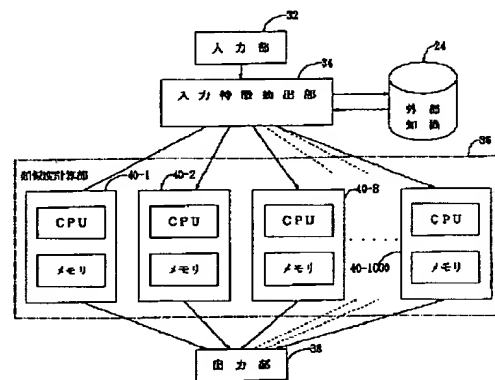
(54) **SIMILARITY DEGREE CALCULATING DEVICE**

(57) Abstract:

PURPOSE: To quickly calculate the degree of similarity between input features acquired from external knowledge and comparison features corresponding to input information and comparison information.

CONSTITUTION: This device is provided with an input part 32 for inputting the input information, the external knowledge 24 for leading features which correspond to supplied information but does not appear in the information itself, similarity degree leading parts 34 and 36 for leading out the features of the input information from the external knowledge 24 and parallelly calculating the degrees of similarity with the respective features of the plural comparison information and an output part 38 for outputting the obtained degrees of similarity.

COPYRIGHT: (C)1994,JPO&Japio



(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平6-274548

(43)公開日 平成6年(1994)9月30日

(51)Int.Cl.⁵

G 0 6 F 15/40
15/38

識別記号

5 1 0 M 9194-5L
C 7323-5L

庁内整理番号

F I

技術表示箇所

審査請求 有 請求項の数 9 O L (全 12 頁)

(21)出願番号 特願平5-61641

(22)出願日 平成5年(1993)3月22日

(71)出願人 000127684

株式会社エイ・ティ・アール自動翻訳電話
研究所

京都府相楽郡精華町大字乾谷小字三平谷5
番地

(72)発明者 大井 耕三

京都府相楽郡精華町大字乾谷小字三平谷5
番地 株式会社エイ・ティ・アール自動翻
訳電話研究所内

(74)代理人 弁理士 深見 久郎 (外2名)

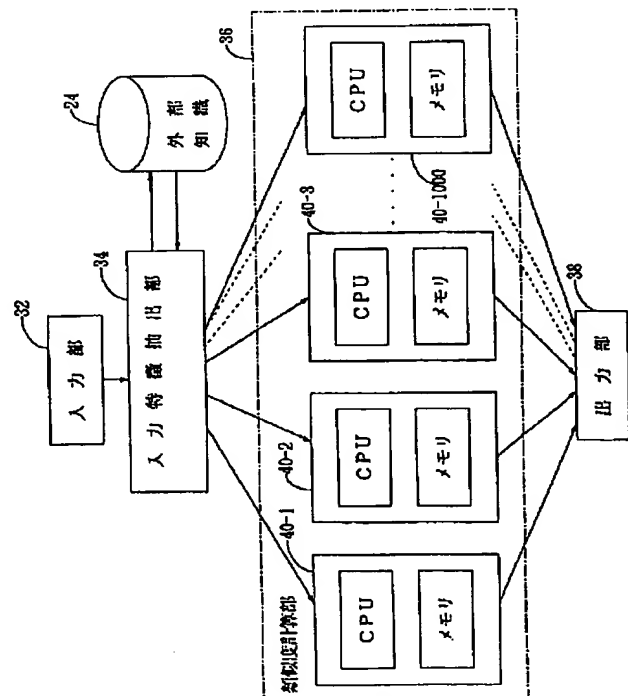
最終頁に続く

(54)【発明の名称】 類似度計算装置

(57)【要約】

【目的】 入力情報と比較情報とに対応して外部知識から獲得した入力特徴および比較特徴の類似度計算を高速化する。

【構成】 入力情報を入力するための入力部32と、与えられる情報に対応するが情報自体には現われていない特徴を導出するための外部知識24と、入力情報の特徴を外部知識24から導出し、複数の比較情報の特徴とのそれぞれと、類似度計算を並列に行なうための類似度導出部34、36と、得られた類似度を出力するための出力部38を含む。



【特許請求の範囲】

【請求項 1】 複数個の予め定める比較情報との間での類似度を計算する対象となる情報を入力するための手段と、

与えられる情報に対応するが、前記与えられる情報自体には現われていない特徴を導出するための外部知識を格納するための手段と、

前記複数個の比較情報に対応して前記外部知識を用いて導出される前記比較情報の特徴のそれぞれと、前記入力手段を介して入力される情報に対応して前記格納手段から導出される特徴との間で所定の類似度計算を並列に行なって、前記比較情報の特徴のそれぞれと、前記入力される情報の特徴との類似度を導出するための手段と、前記導出された類似度を出力するための手段とを含む、類似度計算装置。

【請求項 2】 前記類似度を導出するための手段は、前記複数個の比較情報に対応して前記外部知識を用いて動的に導出される前記比較情報の特徴のそれぞれと、前記入力手段を介して入力される情報に対応して前記格納手段から動的に導出される特徴との間で所定の類似度計算を並列に行なって、各前記比較情報の特徴と前記入力される情報の特徴との間での類似度を導出するための手段を含む、請求項 1 に記載の類似度計算装置。

【請求項 3】 前記外部知識から導出される特徴が n 桁の m 進数で表現されることを特徴とし、 n および m はともに自然数である、請求項 1 に記載の類似度計算装置。

【請求項 4】 前記類似度を導出するための手段が、前記入力される情報について導出される特徴の各桁の文字と、各前記比較情報について導出される特徴の対応する各桁の文字とを比較し、その比較の結果に従って類似度を計算するための手段を含む、請求項 3 に記載の類似度計算装置。

【請求項 5】 前記類似度を計算するための手段が、前記比較の結果相互に一致する文字の数に従って類似度を計算するための手段を含む、請求項 4 に記載の類似度計算装置。

【請求項 6】 前記類似度を導出するための手段が、前記複数個の比較情報のうちの予め定める 1 つに対応して前記外部知識を用いて導出される特徴と、前記入力手段を介して与えられる情報に対応して前記格納手段から導出される特徴との間で所定の類似度計算を行なって、前記複数個の比較情報の前記予め定める 1 つの特徴と前記入力される情報の特徴との間での類似度を導出するための、複数個の、かつ相互に並列に動作可能な演算手段を含む、請求項 1 に記載の類似度計算装置。

【請求項 7】 各前記演算手段が電子計算機を含む、請求項 6 に記載の類似度計算装置。

【請求項 8】 前記複数個の電子計算機が、ネットワークを介して前記入力特徴抽出部と前記出力部とに接続される、請求項 7 に記載の類似度計算装置。

【請求項 9】 前記類似度を導出するための手段が、前記比較情報を格納する、格納内容で参照可能な記憶手段と、

前記入力される情報の特徴の、所定の 1 または複数個の桁の文字と一致する文字を、前記 1 または複数個の桁と一致する位置に有する前記比較情報の特徴を、前記格納内容で参照可能な記憶手段において検出するための手段と、

前記検出手段による検出結果に基づいて各前記比較情報の特徴と前記入力される情報の特徴との間の類似度を計算するための手段を含む、請求項 3 に記載の類似度計算装置。

【発明の詳細な説明】**【0001】**

【産業上の利用分野】 この発明は情報処理分野、特に機械翻訳や情報検索や質問応答などの自然言語処理、画像処理、音声処理などの分野に関し、特に、入力されるある情報（以下「入力情報」と呼ぶ）と、この入力情報との比較が行なわれる他の情報（以下「比較情報」と呼ぶ）のそれぞれとの間の類似度を計算する装置に関する。

【0002】

【従来の技術】 従来、入力情報と比較情報のそれぞれとの間の類似度の計算を行なう類似度計算装置では、計算は入力情報および比較情報自体をデータとして行なわれていた。

【0003】 たとえば、画像検索について述べる。画像検索とは、比較情報として予め準備されている大量の画像の中から、入力画像に類似した画像を検索する処理である。この場合、1 つの入力画像と 1 つの比較画像との間の類似度を求めるときには、各画像の各ドットの色および濃淡のデータ、すなわち、画像そのもののデータを用いていた。

【0004】 これと対照的に、入力情報には現われていない特徴を、外部知識から獲得して類似度計算を行なう事例もある。以下、特に機械翻訳の事例について説明する。

【0005】 機械翻訳システムの 1 タイプとして、大量の用例（原文と訳文との対）と入力文とを比較し、入力文に最も類似した用例を求めてそれを基に訳文を生成するタイプのものがある。そのようなタイプの機械翻訳システムにおいては、入力文すなわち入力情報と用例すなわち比較情報との間の類似度は、その入力文における単語（以下「入力単語」と称する）と用例における単語（以下「比較単語」と称する）との間の類似度を、たとえば類語辞書のような外部知識から獲得した情報を用いて計算する。この類語辞書のような外部知識から獲得した情報は、入力単語あるいは比較単語にそれぞれ対応したものであるが、入力単語あるいは比較単語には現われていないような特徴に関する情報である。

【0006】このように、外部知識から獲得した特徴に関する情報を用いて類似度計算を行なう装置は、従来、入力情報と比較情報のそれぞれとの間の類似度計算を逐次処理で行なっていた。

【0007】図12は、従来のこのような類似度計算装置において、入力情報と比較情報のそれぞれとの間の類似度計算をする際の手順を示すフローチャートである。

【0008】たとえば、電子計算機を使用して類似度計算を行なう場合、比較情報に対応し比較情報には現われていない比較特徴は予め計算機のメモリ上に格納されている。そして類似度計算は次のような手順によって行なわれている。

【0009】図12を参照して、まずステップS001で、入力情報に対応し入力情報には現われていない入力特徴を外部知識から抽出する処理が行なわれる。

【0010】続いてステップS002～S005で、予め計算機のメモリ上に格納されていた比較特徴を1つずつCPU（中央演算処理装置）内に取り出して、順次入力情報の特徴との間の類似度計算を行なっていた。

【0011】

【発明が解決しようとする課題】しかし、入力情報と比較情報とに対応し、入力情報あるいは比較情報には現われていない特徴を用いて類似度計算をする場合には、処理時間が非常にかかるという問題点がある。たとえば、入力文と大量の用例とを比較し、最も類似した用例を求めてそれを基に訳文を生成するタイプの機械翻訳システムにおいて、より正確な翻訳を行なおうとすれば用例数をより多くすることが必要である。ところが、数万の用例を用いて類似度計算を行なう場合には、短い文でも入力文とすべての用例との間の類似度の計算に数分を要する。そのためにこのようなタイプの機械翻訳システムでリアルタイム処理を実現することは不可能であった。

【0012】それゆえに請求項1～9に記載の発明の目的は、入力情報と比較情報のそれぞれとの間の類似度を、外部知識から獲得した、入力情報に対応した入力特徴および比較情報に対応した比較特徴を用いて求める類似度計算において、計算速度を大幅に向上できる類似度計算装置を提供することである。

【0013】

【課題を解決するための手段】請求項1に記載の類似度計算装置は、図1に示されるように、複数の予め定める比較情報との間での類似度を計算する対象となる情報を入力するための手段20と、与えられる情報に対応するが、前記与えられる情報自体には現われていない特徴を導出するための外部知識を格納するための手段24と、前記複数の比較情報に対応して前記外部知識を用いて導出される前記比較情報の特徴のそれぞれと、前記入力手段を介して入力される情報に対応して前記格納手段から導出される特徴との間で所定の類似度計算を並列に行なって、前記比較情報の特徴のそれぞれと、前記入

力される情報の特徴との類似度を導出するための手段22と、前記導出された類似度を出力するための手段26とを含むことを特徴とする。

【0014】請求項2に記載の類似度計算装置は請求項1に記載の装置であって、類似度を導出するための手段は、複数の比較情報に対応して外部知識を用いて動的に導出される比較情報の特徴のそれぞれと、入力手段を介して入力される情報に対応して格納手段から動的に導出される特徴との間で所定の類似度計算を並列に行なって、各比較情報の特徴と入力される情報の特徴との間での類似度を導出するための手段を含む。

【0015】請求項3に記載の類似度計算装置は請求項1に記載の装置であって、外部知識から導出される特徴が n 桁の m 進数（ n および m はともに自然数）で表現されることを特徴とする。

【0016】請求項4に記載の類似度計算装置は請求項3に記載の装置であって、類似度を導出するための手段が、入力される情報について導出される特徴の各桁の文字と、各比較情報について導出される特徴の対応する各桁の文字とを比較し、その比較の結果に従って類似度を計算するための手段を含むことを特徴とする。

【0017】請求項5に記載の類似度計算装置は請求項4に記載の装置であって、類似度を計算するための手段が、比較の結果相互に一致する文字の数に従って類似度を計算するための手段を含むことを特徴とする。

【0018】請求項6に記載の類似度計算装置は請求項1に記載の装置であって、類似度を導出するための手段が、複数の比較情報のうちの予め定める1つに対応して外部知識を用いて導出される特徴と、入力手段を介して与えられる情報に対応して格納手段から導出される特徴との間で所定の類似度計算を行なって、複数の比較情報の予め定める1つの特徴と、入力される情報の特徴との間での類似度を導出するための、複数の、かつ相互に並列に動作可能な演算手段を含む。

【0019】請求項7に記載の類似度計算装置は請求項6に記載の装置であって、各演算手段は電子計算機を含む。

【0020】請求項8に記載の類似度計算装置は請求項7に記載の装置であって、複数の電子計算機が、ネットワークを介して入力手段と外部知識を格納するための手段と出力部とに接続される。

【0021】請求項9に記載の類似度計算装置は請求項3に記載の装置であって、類似度を導出するための手段が、比較情報を格納する、格納内容で参照可能な記憶手段と、入力される情報の特徴の、所定の1または複数の桁の文字と一致する文字を、1または複数の桁の桁と一致する位置に有する比較情報の特徴を、格納内容で参照可能な記憶手段内において検出するための手段と、検出手段による検出結果に基づいて各比較情報の特徴と入力される情報の特徴との間の類似度を計算するための手段

とを含む。

【0022】

【作用】請求項1～9に記載の類似度計算装置においては、入力情報と比較情報のそれぞれとの間の類似度が、外部知識から獲得した、入力情報に対応した入力特徴および比較情報に対応した比較特徴を用いて、並列に計算される。したがって従来の逐次計算処理よりも高速に類似度計算を行なうことができる。

【0023】並列演算を実現するためには、請求項6、7に記載のように、複数個の、相互に並列に動作可能な電子計算機などの演算手段を用いることができる。これら電子計算機は、請求項8に記載のようにネットワークを介して接続されてもよい。また請求項9に記載のように、格納内容で参照可能な記憶手段を用いて類似度の計算を並列的に実行してもよい。

【0024】

【実施例】以下、本発明の3つの実施例を順に説明する。3つの実施例とも、1つの入力情報と1000個の比較情報との間の類似度計算を行なう装置である。情報としては単語を用い、入力された単語（以下「入力単語」と称する）と、比較対象となる単語（以下「比較単語」と称する）1000語との間の類似度計算を行なう例を説明する。

【0025】なお、以下の説明では単語の間の類似度計算を行なう場合を例として本発明を説明するが、本発明は単語の比較のみに限定されるわけではなく、前述のような画像情報や、音声情報などにおける類似度計算に対しても適用可能である。

【0026】実施例の説明をする前に、これら実施例が使用される背景について説明する。前述の、入力文と大量の用例（原文と訳文との対）とを比較して最も類似した用例を求めてそれを基に訳文を生成するタイプの機械翻訳システムについて考える。このタイプの機械翻訳システムにおいては、入力文および用例を構成する各単語ごとの類似度を計算することによって最も類似した用例が求められる。そしてその用例を基に訳文が生成される。

【0027】単語の類似度は以下のようにして求められる。まず、外部知識としての類語辞書を準備する。この類語辞書から、入力単語または比較単語に対応した特徴に関する情報として「類語コード」を獲得し、その類似コードを用いて単語間の類似度を求めている。

【0028】類語辞書とは、単語の意味を階層的に分類した体系に基づき、各単語に類語コードを付与したものを格納した辞書である。図2に、4階層からなる類語辞書の分類体系の一部を示す。

【0029】図2を参照して、各単語に付与される類語コードは3桁の10進数からなる。類語コードの100の位、10の位、1の位はそれぞれ、分類体系の大分類、中分類、小分類を表わす。図2に示される例におい

て、単語「取材」は、大分類が[取引]、中分類が[報道]、小分類が[編集]である分類に属し、その分類の類語コードは457となっている。

【0030】なお、本実施例の説明においては、図2に示すように3桁の10進数の類語コードを用いている。しかし、本発明はこのような類語コードではなく、一般的なn桁m進数やベクトルなど、様々な形の情報を類語コードと同様のものとして処理可能である。

【0031】(1) 第1の実施例

図3を参照して、本発明の第1の実施例の類似度計算装置は、入力部32と、入力特徴抽出部34と、外部知識24と、類似度計算部36と、出力部38とを含む。

【0032】入力部32はキーボード、文字認識装置、音声認識装置などからなる。入力部32は、入力単語を入力特徴抽出部34に与えるためのものである。

【0033】外部知識24は、ハードディスクやメモリなどからなる。この外部知識24は、図4に示されるような、各単語の見出しとその類語コードが対になったデータを多数格納している。

【0034】入力特徴抽出部34は、入力部32から与えられる入力単語に対応した類語コード（以下「入力コード」と称する）を外部知識24から抽出し、類似度計算部36に与えるためのものである。

【0035】類似度計算部36は、1000個のコンピュータ40-1～40-1000を含む。各コンピュータ40-1～40-1000の入力は入力特徴抽出部34の出力に接続されている。また各コンピュータ40-1～40-1000の出力は出力部38の入力に接続されている。これらコンピュータ40-1～40-1000の各々は、CPUとメモリとを含む。1つのコンピュータのメモリには、1つの比較単語に対応した類語コード（以下「比較コード」と称する）および類似度計算のためのプログラムが格納されている。これらコンピュータ40-1～40-1000には入力特徴抽出部34から入力コードが与えられる。各コンピュータ40-1～40-1000は、与えられた入力コードと各コンピュータのメモリに格納されている比較コードとの間の類似度を後述するような方法に従って計算し、計算された類似度を出力部38に与える。これらコンピュータ40-1～40-1000の各メモリに格納されている比較コードの一例が図5に示されている。図5においてたとえばコンピュータ40-1に格納されている比較コード「426」は、図4に示されるように比較単語「販売」に対応する類語コードである。他の比較コードも同様に図4に示される外部知識24内の或る単語に対応する類語コードとなっている。

【0036】出力部38は、表示装置、印刷装置などからなる。出力部38は、類似度計算部36のコンピュータ40-1～40-1000から与えられた類似度を出力するためのものである。

【0037】図6は、図3のコンピュータ40-1~40-1000の各々で行なわれる、入力コードと比較コードとの間の類似度の求め方を示す。図6において最左欄入力コードと比較コードとの間に成立する条件を示す。中欄は、最左欄に示される条件に適合したときの類似度を示す。最右欄は、最左欄に示される条件に適合するような入力コードと比較コードとの対を示す。

【0038】入力コードと比較コードとの条件欄における記号「I1」「I2」「I3」はそれぞれ、入力コードの100の位と、10の位と、1の位とを表わす。記号「C1」「C2」「C3」はそれぞれ、比較コードの100の位と、10の位と、1の位とを表わす。

【0039】図6を参照して、第2行目で示されるように、入力コードと比較コードとが、100の位と、10の位と、1の位とのいずれでも一致する場合には、類似度は3となる。第3行目に示されるように100の位と10の位とが一致し、1の位のみが異なる場合には類似度は2となる。100の位のみが一致し、10の位が異なる場合には、図6の4行目に示されるように類似度は1となる。図6の第5行目に示されるように入力コードと比較コードの100の位が互いに異なる場合には類似度は0となる。

【0040】以下、図3に示される類似度計算装置の動作を、入力単語が「取材」である場合を例にして説明する。入力単語「取材」が入力部32を介して入力されると、その入力単語「取材」は入力特徴抽出部34に与えられる。

【0041】入力特徴抽出部34は、入力単語「取材」に対応した類語コード（入力コード）を外部知識24から抽出する。この場合図4の第4行目に示されるように、入力単語「取材」に対応した類語コードは457となっているので、入力コードとして457が得られる。この入力コード457は、類似度計算部36のコンピュータ40-1~40-1000のすべてに与えられる。

【0042】各コンピュータ40-1~40-1000は、入力コード「457」が与えられると、それぞれ独立に類似度計算を行なう。各コンピュータは、入力コード「457」と、そのコンピュータに割り当てられている比較コードとを比較し、図6に示される類似度の求め方に従って類似度を求める。

【0043】たとえば図5を参照して、コンピュータ40-1のメモリに格納されている比較コードは「426」である。したがって図6の第4行目の条件（ハ）により類似度1となる。コンピュータ40-2では、比較コードは「149」である。条件（ニ）により類似度0となる。コンピュータ40-3では、比較コードは「458」である。条件（ロ）により類似度は2となる。コンピュータ40-1000では、比較コードは「732」である。条件（ニ）により類似度0となる。他のコンピュータ40-4~40-999でも同様の類似度計

算が行なわれる。類似度計算部36の各コンピュータ40-1~40-1000は、求められた類似度を出力部38に与える。

【0044】出力部38は、コンピュータ40-1~40-1000からの類似度がすべて与えられると、表示装置や印刷装置などにその類似度を出力する。

【0045】この第1の実施例では、1000個の類似度計算がコンピュータ40-1~40-1000により並列に行なわれる。したがって従来の逐次処理による計算に比べて単純計算で約1000倍高速に行なわれる。

【0046】（2） 第2の実施例

図7は、本発明の第2の実施例の類似度計算装置の概略構成図である。図7を参照してこの類似度計算装置は、入出力管理部50と、入出力管理部50が接続されるネットワーク52と、ネットワーク52に接続される1000個の類似度計算部54-1~54-1000とを含む。

【0047】入出力管理部50は、入力部32と、外部知識24と、入力特徴抽出部34と、出力部38とを含む。図7と図3とにおいて、同一のブロックには同一の参照符号および名称が与えられており、それらの機能も同一である。したがってここではそれらについての詳しい説明は繰返さない。この入出力管理部50は、ワークステーションやパーソナルコンピュータなどによって構成される。

【0048】外部知識24には、図4に示されるデータが格納されている。類似度計算部54-1~54-1000の各々は、ワークステーションやパーソナルコンピュータなどを含む。これら各類似度計算部54-1~54-1000は、CPUとメモリとを含む。1つの類似度計算部のメモリには、1つの比較単語に対応した類語コード（比較コード）および類似度計算のプログラムが格納されている。

【0049】以下、この発明の第2の実施例の類似度計算装置の動作を説明する。キーボードなどからなる入力部32によって入力単語が入力特徴抽出部34に与えられる。入力特徴抽出部34は、この入力単語に対応する入力コードを外部知識24から抽出し、ネットワーク52を介して類似度計算部54-1~54-1000のすべてに送る。各類似度計算部54-1~54-1000は、送られてきた入力コードと、各類似度計算部のメモリに格納されている比較コードとの間の類似度を第1の実施例における方法と同様の手順で計算する。類似度計算部54-1~54-1000はすべて、求めた類似度を入出力管理部50にネットワーク52を介して送る。入出力管理部50の出力部38は、前述と同様に表示装置、印刷装置などからなり、類似度計算部54-1~54-1000から送られてきた類似度を出力する。

【0050】この第2の実施例においても、類似度計算部54-1~54-1000はそれぞれ独立に類似度計

算を行なう。すなわち、1000個の類似度計算が並列に行なわれる。したがって従来の逐次処理による計算に比べて単純計算で約1000倍高速に類似度計算が行なわれる。

【0051】(3) 第3の実施例

図8は、本発明の第3の実施例の類似度計算装置の概略構成を示すブロック図である。図8を参照して、この類似度計算装置は、入力部32と、外部知識24と、入力特徴抽出部34と、類似度計算部60と、出力部38とを含む。図8と図3とにおいて、同一のブロックには同一の参照符号および名称が与えられている。それらの機能も同一である。したがってここではそれらについての詳しい説明は繰返さない。

【0052】類似度計算部60は、CPU62と、メモリ64と、内容でアドレス可能な連想メモリ66とを含む。

【0053】メモリ64には、類似度計算のプログラムが格納されている。またメモリ64には、各比較情報に対応する類似度を格納するエリアが設けられている。

【0054】連想メモリ66の各ワードには、1つの比較単語に対応した類語コード（比較コード）が予め格納されている。そして合計1000個のワードに1000個の比較コードが格納されている。

【0055】連想メモリ66は、マスクによる一致検索機能と部分並列書込機能とを有するものとする。マスクによる一致検索機能とは、ビット列の特定部分に対して一致検索を行なう機能をいう。部分並列書込機能とは、一致が検出された複数のデータの、特定のビットに並列にデータを書込む機能をいう。

【0056】連想メモリ66のデータの格納形式が図9に示される。図9を参照して、連想メモリ66には比較コード1～1000の1000個のエリアが設けられる。各比較コードのためのエリア、たとえば比較コード1のエリアは全部で15ビットからなる。このうち先頭から3ビットは類似度を求めるために使用されるエリアである。後半の12ビットは比較コードを格納するためのエリアである。

【0057】類似度を求めるためのエリアの3ビットの第1のビットは、入力コードと比較コードとの100の位の一致／不一致を示すビットであり、2ビット目は入力コードと比較コードとの10の位の一致／不一致を示すビットであり、3ビット目は入力コードと比較コードとの1の位の一致／不一致を示すビットである。

【0058】比較コードを示す12ビットはそれぞれ4ビットずつの3つの領域に分けられる。これら3つの領域は、図9に示されるように順に比較コードの100の位を表わす4ビットと、比較コードの10の位を表わす4ビットと、比較コードの1の位を表わす4ビットとである。

【0059】なお、各比較コードエリアのうち最初の3

ビットは、類似度計算に先立って初期値たとえばすべて0に設定されているものとする。

【0060】以下、この第3の実施例の類似度計算装置の動作を説明する。キーボード、文字認識装置、音声認識装置などからなる入力部32により入力単語が入力されると、その入力単語が入力特徴抽出部34に与えられる。入力特徴抽出部34は、与えられる入力単語に対応した類語コード（入力コード）を外部知識24から抽出する。外部知識24に格納されているデータは図4に示されたものと同様である。

【0061】抽出された入力コードは類似度計算部60に与えられる。類似度計算部60は、次のようにして入力コードと1000個の比較コードとの間の類似度計算を行ない、その結果を出力部38に与える。

【0062】類似度計算部60では、連想メモリ66を用いた次のような類似度計算が行なわれる。類似度の定義は、第1の実施例において図6を参照して説明したものと同じである。

【0063】以下、入力単語として「取材」が入力された場合の、類似度計算部60の動作を説明する。入力単語「取材」に対応する類語コード（入力コード）は図4に示されるように「457」である。この入力コード「457」が類似度計算部60に与えられると、類似度計算部60は次のように動作する。

【0064】まず、図8のメモリ64内に、図11に示すように、図10に示される連想メモリ内の比較コードの格納エリアと同様のデータ構造を有するエリアを設ける。このエリアは8行のエリア70、72、74、76、78、80、82、84からなる。第1行目のエリア70の後半の12ビットには、入力コードの「457」が格納される。このエリア70の先頭の3ビットはすべて0である。この1行目のエリア70を入力を検索コードと呼ぶ。

【0065】第2行目のエリア72には、入力の検索コードのうち、その100の位を表わす4ビット以外をマスクしたデータが格納される。エリア74には、入力の検索コードのうち10の位を表わす4ビット以外をマスクしたデータが格納される。エリア76には、入力の検索コードのうち1の位を表わす4ビット以外をマスクしたデータが格納される。エリア78には、先頭の3ビットに「111」が格納され、それ以外のビットがマスクされたデータが格納される。エリア80には、先頭の3ビットに「110」が格納され、それ以外のビットがマスクされたデータが格納される。エリア82には、先頭の2ビットに「10」が格納され、それ以外のビットがマスクされたデータが格納される。エリア84には、先頭のビットに「0」が格納され、それ以外のビットがすべてマスクされたデータが格納される。以下、次に示すような手順に従って類似度計算が行なわれる。

【0066】① エリア72に格納された、検索コード

の100の位を表わす4ビット以外をマスクしたデータによる一致検索の命令と、一致した比較コードの100の位の一致／不一致を示すビットに1を書込む命令とを順に連想メモリ66に与える。

【0067】② エリア74に格納された、検索コードの10の位を表わす4ビット以外をマスクしたデータによる一致検索の命令と、一致した比較コードの10の位の一致／不一致を示すビットに1を書込む命令とを順に連想メモリ66に与える。

【0068】③ エリア76に格納された、検索コードの1の位を表わす4ビット以外をマスクしたデータによる一致検索の命令と、一致した比較コードの1の位の一致／不一致を示すビットに1を書込む命令とを順に連想メモリ66に与える。

【0069】上述の①～③の処理を行なった結果、連想メモリ66上のデータは図10に示されるようになる。すなわち比較コード1(426)と入力コード「457」とは100の位のみが一致するために、図10の比較コード1のエリアに示されるように100の位の一致／不一致を示すビットのみが“1”となり、他の2ビットは“0”となる。比較コード2(149)と入力コードとは一致する桁がないために比較コード2の先頭の3ビットは“000”となる。同じように比較コード3の先頭の3ビットは“110”となる。

【0070】④ 図11のエリア78に格納されたデータによる一致検索命令を連想メモリ66に与え、一致した比較コードを検出する。そして、図8のメモリ64上に予め準備されていた、各比較情報に対応する類似度を格納するエリアのうち、一致が検出された比較コードに対応するエリアに類似度として「3」を格納する(図6の条件(イ))。

【0071】⑤ 図11のエリア80に格納されたデータによる一致検索命令を連想メモリ66に与え、一致した比較コードを検出する。そして、一致が検出された比較コードに対応するメモリ64上のエリアに類似度として「2」を格納する(図6の条件(ロ))。

【0072】⑥ 図11のエリア82に格納されたデータによる一致検索命令を連想メモリ66に与え、一致した比較コードを検出する。そして、1が検出された比較コードに対応するメモリ64上のエリアに類似度として「1」を格納する(図6の条件(ハ))。

【0073】⑦ 図11のエリア84に格納されたデータによる一致検索命令を連想メモリ66に与え、一致した比較コードを検出する。1が検出された比較コードに対応するメモリ64上のエリアに類似度として「0」を与える。(図6の条件(ニ))。

【0074】⑧ 入力コードとすべての比較コードとの間の類似度を出力部38に与える。上記した①～⑦で行なわれる一致検索命令および書込命令は連想メモリ上のすべての比較コードに対して並列に行なわれる。したが

って第1の実施例と同様に、従来の逐次処理による類似度計算に比べて、単純計算で約1000倍高速となり、類似度計算を高速に実現することができる。

【0075】このように本発明に係る類似度計算装置を用いれば、入力情報と比較情報のそれぞれとの間の類似度を、外部知識から獲得した、入力情報に対応した入力特徴および比較情報に対応した比較特徴を用いて並列に行なわれる類似度計算によって求めることができる。特に、入力文と大量の用例(原文と訳文との対)とを比較して最も類似した用例を求めてそれを基に訳文を生成するタイプの機械翻訳システムや、大量の情報と入力情報との間の類似度を求める必要がある画像検索や文章検索などの情報検索において、入力情報に対応した特徴を用いて類似度を求める必要がある場合に、高速にそうした処理を行なうことが可能となる。

【0076】

【発明の効果】以上のように請求項1ないし9に記載の類似度計算装置によれば、複数の比較情報に対応して外部知識を用いて導出される比較情報の特徴のそれぞれと、入力情報に対応して導出される特徴との間で所定の類似度計算が並列に行なわれて類似度が導出される。したがって、このようにして比較特徴を用いて求める類似度計算において、従来の逐次処理と異なって計算速度の大幅な向上を実現できる類似度計算装置を提供できる。

【図面の簡単な説明】

【図1】本発明に係る類似度計算装置の概略の構成を示す図である。

【図2】類語辞書の分類体系の一部を示す図である。

【図3】本発明の第1の実施例の類似度計算装置の概略構成を示すブロック図である。

【図4】外部知識の内容を示す模式図である。

【図5】この発明の第1の実施例の類似度計算装置の類似度計算部内の計算機メモリに格納されている比較コードを示す図である。

【図6】この発明における入力コードと比較コードとの間の類似度の求め方を示す図である。

【図7】本発明の第2の実施例の類似度計算装置の概略構成を示すブロック図である。

【図8】本発明の第3の実施例の類似度計算装置の概略構成を示すブロック図である。

【図9】この発明の第3の実施例の類似度計算装置における連想メモリ上の比較コードの格納形式を示す図である。

【図10】この発明の第3の実施例における連想メモリ上の比較コードのデータを示す図である。

【図11】この発明の第3の実施例の類似度計算装置の行なう類似度計算において使用する、メモリ上の検索のためのデータを示す図である。

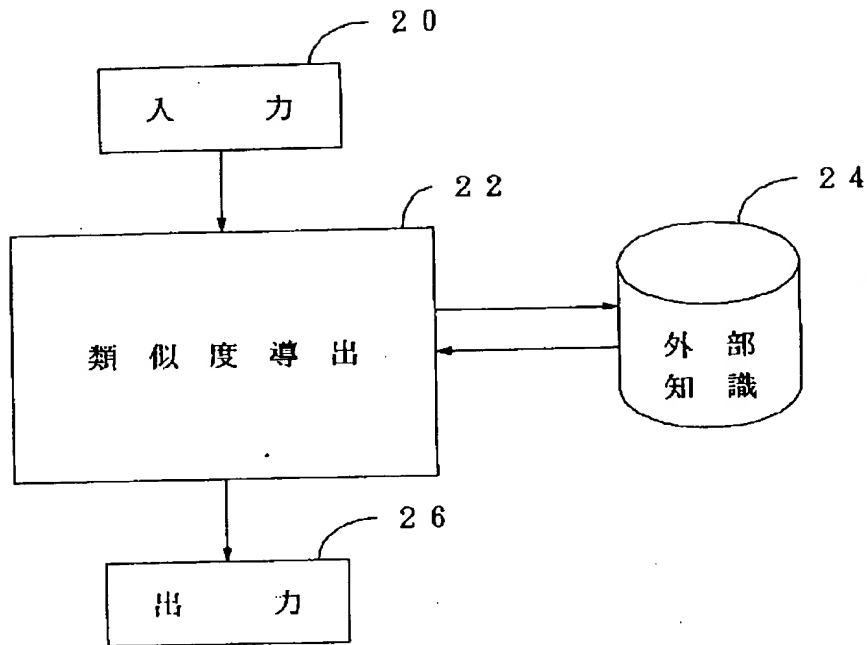
【図12】従来の類似度計算装置における類似度計算の手順を示す図である。

【符号の説明】

24 外部知識
32 入力部
34 入力特徴抽出部
36 類似度計算部

38 出力部
50 入出力管理部
52 ネットワーク
60 類似度計算部
66 連想メモリ

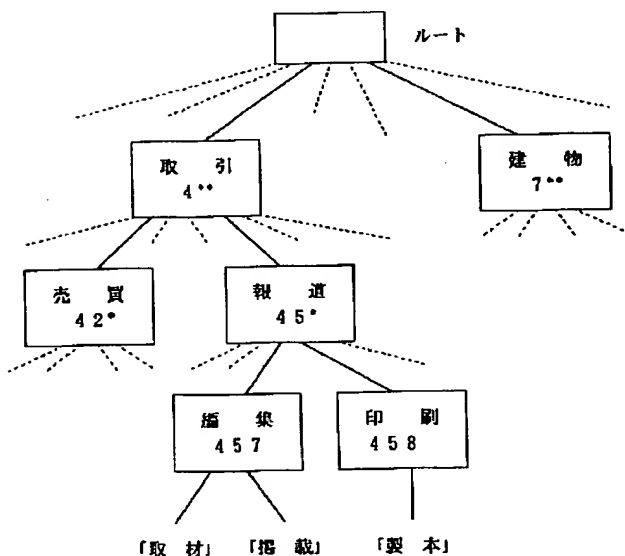
【図1】



【図4】

単語の見出し	類似コード
製 本	4 5 8
建 物	7 3 2
取 材	4 5 7
白 鳥	1 4 9
販 売	4 2 6
掲 載	4 5 7
⋮	⋮

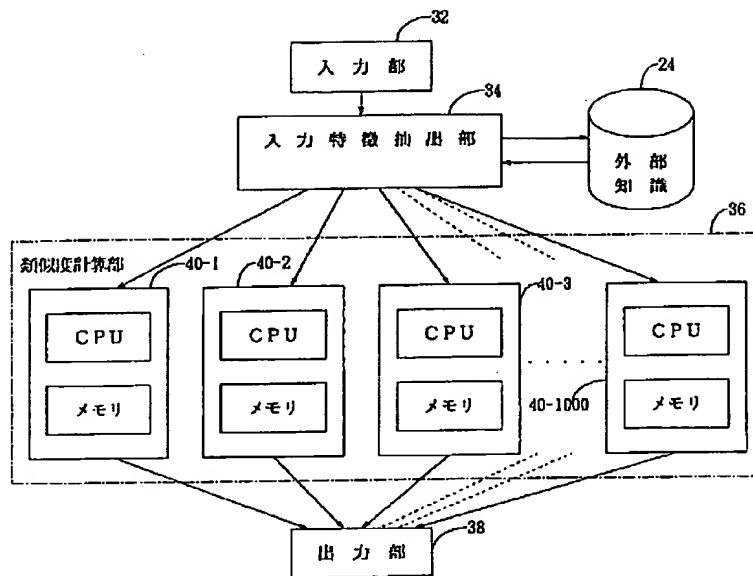
【図2】



【図6】

入力コードと比較コードの条件		類似度	例 (入力コード: 比較コード)
(イ)	$I_1, I_2, I_3 = C_1, C_2, C_3$	3	4 5 7 : 4 5 7
(ロ)	$I_1, I_2 = C_1, C_2, I_3 \neq C_3$	2	4 5 7 : 4 5 8
(ハ)	$I_1 = C_1, I_2 \neq C_2$	1	4 5 7 : 4 2 6
(ニ)	$I_1 \neq C_1$	0	4 5 7 : 1 4 9

【図3】



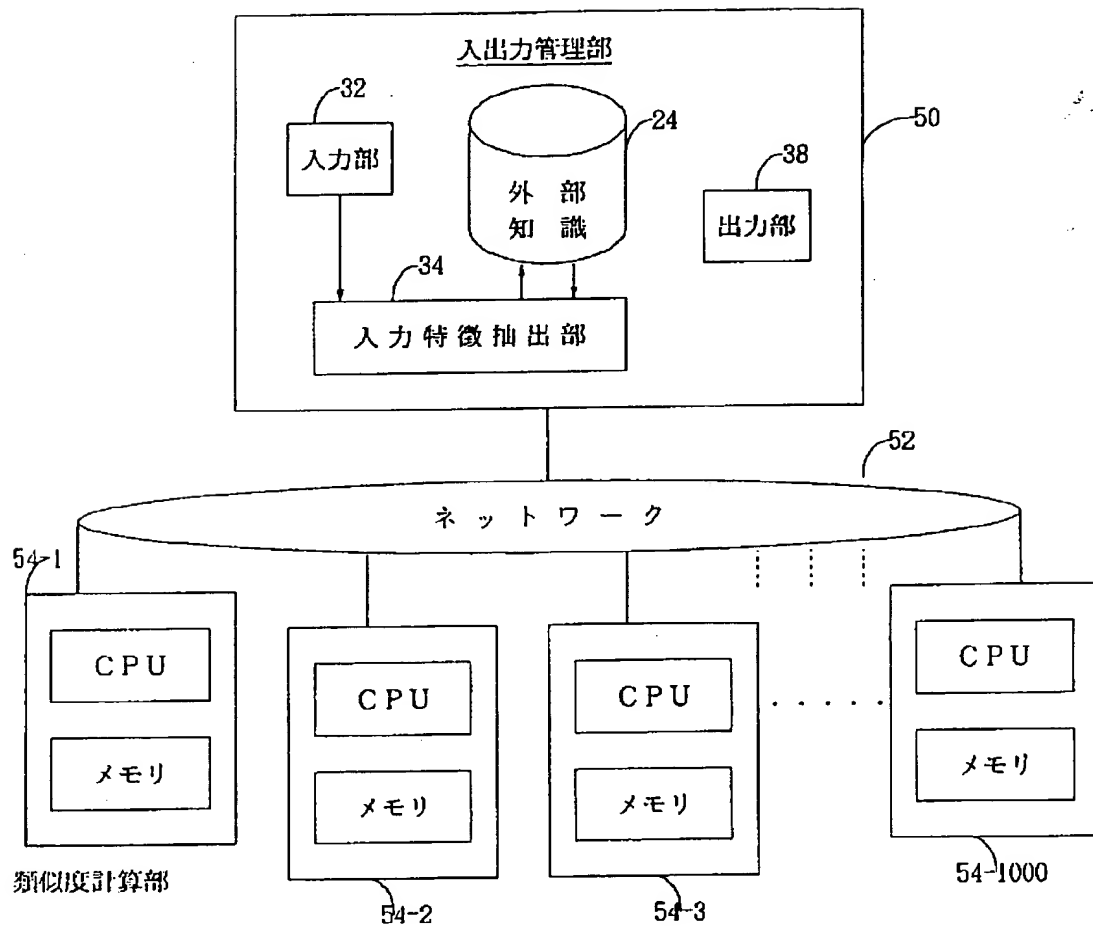
【図5】

類似度計算部コボータ	比較コード	
40-1	426	← 比較単語「販売」に対応する類語コード
40-2	149	← 比較単語「白鳥」に対応する類語コード
40-3	458	← 比較単語「製本」に対応する類語コード
⋮	⋮	
40-1000	732	← 比較単語「建物」に対応する類語コード

【図9】

	百の位の一致/不一致			十の位の一致/不一致			一の位の一致/不一致		
	コードの百の位			コードの十の位			コードの一の位		
比較コード1	0	0	0	0	1	0	0	0	1
比較コード2	0	0	0	0	0	0	1	0	0
比較コード3	0	0	0	0	1	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
比較コード1000	0	0	0	0	1	1	0	0	1

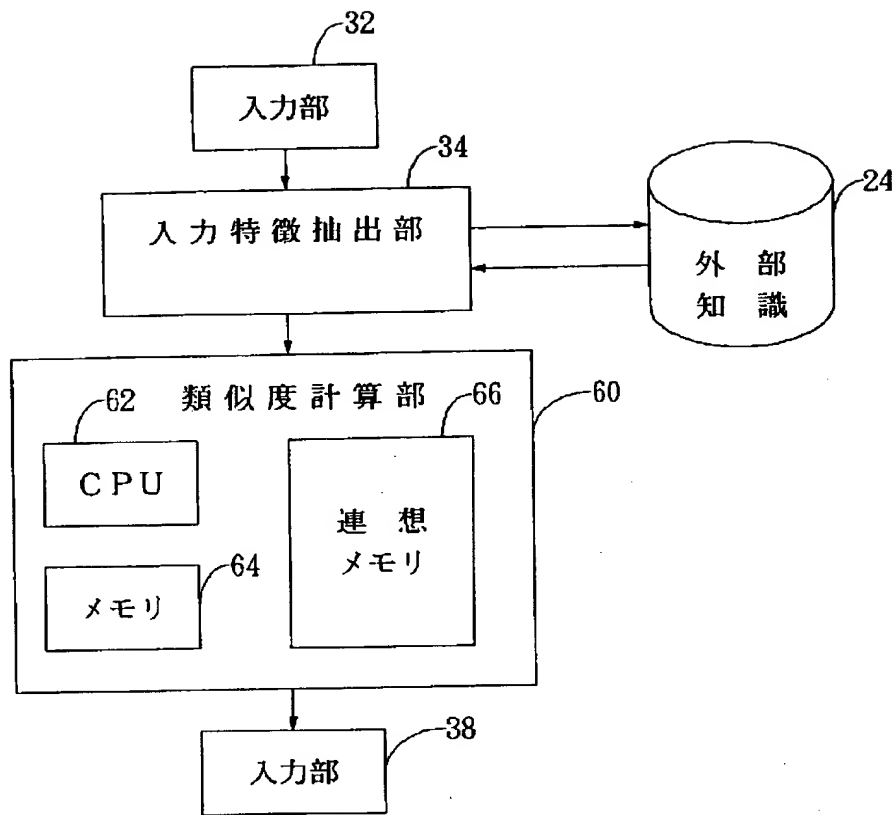
【図7】



【図10】

	百の位の一致/不一致			十の位の一致/不一致		一の位の一致/不一致	
	コードの百の位			コードの十の位		コードの一の位	
比較コード1	1	0	0	0	1	0	0
比較コード2	0	0	0	0	0	0	1
比較コード3	1	1	0	0	1	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
比較コード1000	0	0	0	0	1	1	0

【図8】



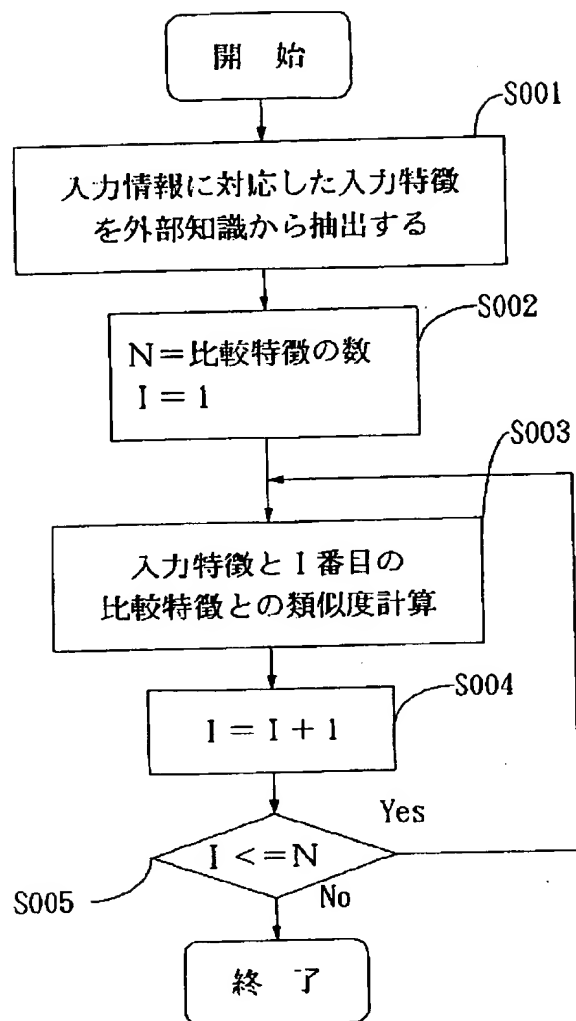
【図11】

70: 入力の検索コード

	0	0	0	0100	0101	0111
72~	X	X	X	0100	XXXX	XXXX
74~	X	X	X	XXXX	0101	XXXX
76~	X	X	X	XXXX	XXXX	0111
78~	1	1	1	XXXX	XXXX	XXXX
80~	1	1	0	XXXX	XXXX	XXXX
82~	1	0	X	XXXX	XXXX	XXXX
84~	0	X	X	XXXX	XXXX	XXXX

百の位的一致/不一致
十の位的一致/不一致
一の位的一致/不一致
コードの百の位
コードの十の位
コードの一の位

【図12】



フロントページの続き

(72)発明者 隅田 英一郎
京都府相楽郡精華町大字乾谷小字三平谷5
番地 株式会社エイ・ティ・アール自動翻
訳電話研究所内

(72)発明者 飯田 仁
京都府相楽郡精華町大字乾谷小字三平谷5
番地 株式会社エイ・ティ・アール自動翻
訳電話研究所内